Research Report

One Approach to Detecting the Invariance of Proficiency Standards Over Time

Jiahe Qian

April 2008

ETS RR-08-15

One Approach to Detecting the Invariance of Proficiency Standards Over Time

Jiahe Qian
ETS, Princeton, NJ

April 2008

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

This study explores the use of a mapping technique to test the invariance of proficiency standards over time for state performance tests. First, the state proficiency standards are mapped onto the National Assessment of Educational Progress (NAEP) scale. Then, rather than looking at whether there is a deviation in proficiency standards directly, the invariance of their NAEP equivalents is tested over time. The basis of the mapping technique is an enhanced method that was originally designed for comparing performance standards for public school students set by different states when the state tests are comparable. This approach can also be used to detect score inflation over time for state tests.

Key words: Proficiency standards, proficiency standard deviation, score inflation, equipercentile linking, NAEP equivalents, No Child Left Behind

Acknowledgments

The author thanks Shelby Haberman, Henry Braun, Sandip Sinharay, and William Monaghan for their suggestions and comments. The author is particularly grateful to Daniel Eignor for his help in clarifying the psychometric meaning of the statistical techniques. The author would also like to thank Bruce Kaplan, Sailesh Vezzu, and Xiaoke Bi for their computational assistance and Kim Fryer for editorial assistance. The opinions expressed herein are solely those of the author and do not necessarily represent those of Educational Testing Service. This work was supported by the National Center for Education Statistics, Contract # ED-02-CO-0023.

Table of Contents

List of Tables

Page

List of Figures

Page

1. Introduction

Recently the stability of state performance test standards has been a concern in education because under the No Child Left Behind Act (NCLB), each state can select its own tests and set its own proficiency standards for reading and mathematics to determine its standing with respect to the national requirements of adequate yearly progress (American Federation of Teachers, 2006). This study was designed to develop an approach to test the invariance of proficiency standards over time for state tests or analogous assessments. Proficiency standards, specific levels of mastery of knowledge and skills in education, are usually anchored by cut points on a test scale; the cut points classify student performance into several achievement categories, such as *basic*, *proficient*, or *advanced*. State tests usually use the equating process to maintain the numerical cut points related to proficiency standards, provided that no substantial changes occur in assessment. But over time the proficiency standards could deviate from the achievement levels on the scale on which they were established, so that each cut point no longer anchors the same ability level. This phenomenon is called *deviation in proficiency standards* (DPS) or *proficiency standard deviation.* Many researchers have found DPS in performance assessments when they are compared with other stable assessments, such as the National Assessment of Educational Progress (NAEP; Cannell, 1987; Grissmer, Flanagan, Kawata, & Williamson, 2000; Klein, Hamilton, McCaffrey, & Stecher, 2000; Neill et al., 1997; Smith, 1991). Note that another concept, *scale drift*, is also related to the stability of a test scale, but, as Angoff (1984) noted, scale drift is usually related to less than adequate equating of new form of a test to "one or more of the existing forms for which conversions to the reference scale (i.e., the reporting scale) are already available" (p. viii).

DPS could be caused by many factors, such as score inflation, scale drift, alteration of the test instrument, changes in assessment format, reform of the subject framework, content modification, or differential performance gains (Koretz, 2007; Madaus, 1988b). However, if other factors are not present, DPS can serve as an indicator of score inflation, meaning that students are getting higher test scores than before at each given level of academic achievement (Arenson, 2004; Koretz, 1988; Linn, 2000; Potter, 1979). As an important application of testing DPS, section 4 will discuss the procedure of detecting score inflation.

The approach to testing the invariance of proficiency standards is based on an enhanced mapping method, which was originally designed for comparing performance standards set for

public school students by different states when the tests are comparable (Braun & Qian, 2007a). It is hard to test whether the proficiency standard deviates from its original scale by simply observing the changes of scores on a test itself, but a test that potentially has DPS can be compared with another test that has no DPS. For example, many educators have compared state tests with the NAEP assessments. They have found that test score improvements shown on state tests used for high-stakes decisions may not be corroborated by score improvements on NAEP (Haney, 2002; Linn, Graue, & Sanders, 1990). This strategy will be to measure the invariance of proficiency standards by transforming the scale of the state test to the well established NAEP scale and use the NAEP scale as a benchmark for comparison.

To implement this approach, the state proficiency standards are first mapped onto the NAEP scale. These mapped proficiency standards of state tests are called the *NAEP equivalents to the state standards* or *NAEP equivalents.* Then, rather than directly detecting invariance of proficiency standards, this paper looks at the related solution: the invariance of the NAEP equivalents over time. The mapping makes the comparison effective because, as a benchmark, NAEP is generally regarded as meeting high standards with respect to test design, test content, and psychometric quality. In addition, NAEP is the only nationally standardized test that is administered in a uniform and stable manner across states. Also, NAEP scores are not influenced by factors such as grade inflation. For a general introduction to NAEP, see Jones & Olkin (2004). Although literature demonstrates that for a number of reasons, linking state tests to NAEP assessments at the student level does not result in an appropriate or valid linking (Feuer, Holland, Green, Bertenthal, & Hemphill, 1998; Koretz, Bertenthal, & Green, 1999), many studies have shown that the mapping of the proficiency standards on state tests to NAEP equivalents is valid (Braun & Qian, 2007b; McLaughlin & Bandeira de Mello, 2003). These studies have further concluded that most of the heterogeneity across states in the NAEP equivalents to the state standards can be attributed to differences in the stringency of proficiency standards set by the states.

This study assumes that two assessments of a given subject, one state test and one NAEP assessment, are reasonably equivalent. It takes for granted that a state test appropriately maintains its numerical cut points related to the proficiency standards over time via the equating process. Moreover, it is assumed that both the state test and the NAEP assessment maintain their own testing instruments and other conditions over time.

When a significant change in the NAEP equivalents over time has been detected, it suggests that the proficiency standard of a state test has deviated from its original scale. To confirm the causes of significant DPS, especially to claim score inflation, this paper suggests forming a committee, with members consisting of test experts and subject-matter specialists, to judge causes of the observed change.

Section 2 of this paper will provide a description of the estimation method for mapping proficiency standards of state tests and introduce some properties of the mapped proficiency standards over time. Section 3 introduces the data used in the study, namely the 2003 and 2005 fourth and eighth grade state tests of reading and mathematics, and presents the empirical results from testing the invariance of state standards. Section 4 applies the approach in detecting score inflation for state tests. Section 5 offers a summary and some conclusions.

2. Methodology

Before presenting the approach to detect the invariance of state proficiency standards, the next section introduces the procedure that maps state proficiency standards onto the NAEP scale.

2.1 Outline of the Methodology for Mapping State Standards to NAEP Scale

As described in Braun & Qian (2007a), the mapping procedure was carried out separately for each state that participated in NAEP and was represented in the National Longitudinal School-level State Assessment Score Database[1] (for the corresponding academic year). To make the comparisons of the NAEP equivalents over time effective, both the state tests and NAEP assessments need to comply with standard conditions, which will be introduced in section 2.2. The statistical analysis in this study involves the NAEP sample design, school weights, and target estimation, among other things. In NAEP, state samples are obtained through a two-stage probability sampling design. To account for the unequal probabilities of selection and to allow for adjustments for nonresponse, each school and each student are assigned separate sampling weights. In this study, appropriate weights were applied in the estimation of the proportion of students in the state who scored above the standard. The statewide target proportion of students meeting the standard is estimated by a ratio estimator. Appendix A in this paper provides a description of the weights, target estimation, and variance estimation.

Let P denote the state-wide proportion of students meeting a particular standard. Let F denote the score distribution on the NAEP assessment for the state and the $(1-P)$th quantile on

F be $\xi = F^{-1}(1-P)$. The estimate of the $(1-P)$th quantile, $\hat{\xi}$, can also be denoted as $\hat{\xi}_{WAM}$ where the abbreviation WAM stands for "weighted aggregate mapping". Braun & Qian (2007a) followed the steps below in the procedure for mapping state standards to NAEP scale:

1. Based on the proportions of students who meet a given state's performance standard on that state's own assessment in NAEP-sampled schools, estimate the proportion of students in the state as a whole who meet the state's standard.

   First, the schools in the state NAEP sample are identified and matched with their records in the National Longitudinal School-Level State Assessment Score Database. For each school, the proportion of students meeting the state standard is obtained. By using the school weights from the NAEP design, an estimate is obtained of P using a ratio estimator, $\overline{p}_w$, which is a weighted average estimate of the number of students meeting the standard over a weighted average estimate of the number of eligible students. For a more detailed description of the weights and the ratio estimator, see Appendix A.

2. Based on the NAEP sample of schools and students within schools, estimate the distribution of scores on the NAEP assessment for the state as a whole.

   This procedure is carried out to generate the results contained in the report that is issued after each NAEP assessment. Let $\hat{F}$ denote the empirical distribution of F, which can be obtained based on the NAEP sample.

3. Find the point on the NAEP score scale at which the estimated proportion of students in the state scoring above that point equals the proportion of students in the state meeting the state's own performance standard.

   After the proportion P of students meeting the state's own performance standard (defined with respect to the state test score scale) is estimated by $\overline{p}_w$ and the NAEP score distribution is calculated as in Steps 1 and 2, the performance standard is mapped to the NAEP scale by finding the point $\hat{\xi}$ on the NAEP scale that is the $(1-\overline{p}_w)$th quantile:

$$\hat{\xi}_{\mathrm{WAM}} = \hat{F}^{-1}\left(1 - \overline{p}_w\right). \qquad (1)$$

The estimated NAEP equivalent to the state standard is taken to be $\hat{\xi}_{WAM}$, which is an estimate of ξ. If the state employs more than one standard, this procedure can be repeated for each one.

4. Compute an estimate of the variance of the estimated NAEP equivalent.

   This computation is developed based on the NAEP jackknife methods to obtain variance estimates given NAEP's complex sample design and latent ability measurement (Allen, Donoghue, & Schoeps, 2001).

Figure 1 illustrates the mapping procedure. The dashed curve on the left-hand side represents an estimate of the state distribution of scores on the state test, based on the scores of all students in the schools selected for the state's NAEP sample. The area in the upper tail of this distribution above the state standard is an estimate of the proportion of students in the state meeting or exceeding that standard, and is denoted by $\hat{p}_w$. In practice, it is only necessary to obtain $\hat{p}_w$ from the data. The curve on the right-hand side represents the estimated distribution of NAEP scores for the state. This is the usual reported NAEP distribution that is estimated based on the performance of students in the state's NAEP sample who took the NAEP assessment. The estimated NAEP equivalent to the state standard, $\hat{\xi}$, is the point on the NAEP scale such that the corresponding upper tail area of the NAEP distribution also equals $\hat{p}_w$. For a given distribution of state test scores and a specific distribution of NAEP assessment scores, by the monotone property of equipercentile linking, a larger $\hat{p}_w$ corresponds to a lower $\hat{\xi}$ and vice versa.

2.2. Testing the Invariance of State Standards Over Time

Mapping state standards to NAEP scale over time. As pointed out in the introductory section, the validity of the mapping methodology requires that the state test and the NAEP assessment be reasonably equivalent with respect to their test instruments, including subject frameworks, assessment format, psychometric characteristics of the tests, norms, and so on. Next, the standard conditions involved with the procedure are described. The following is assumed: (a) there are no considerable changes in the state test instrument over time, (b) the state

tests maintain their numerical cut points related to the standards over time (via score equating), and (c) the distributions of state test scores over time maintain the same shape and spread, but there is allowance for horizontal shifting of the distribution curve. The same assumptions are applied to the NAEP assessments. These standard conditions are reasonable even though they may appear to be stringent.

Let z_A and z_B be the state test standards of time point A and time point B respectively. Because state tests are assumed to maintain their standards over time, $z_A = z_B$. Let ξ^A and ξ^B be the images of z_A and z_B for time point A and time point B, respectively. Their estimates are $\hat{\xi}^A$ and $\hat{\xi}^B$. The variance estimation of $\hat{\xi}^A$ or $\hat{\xi}^B$ is the same as that for $\hat{\xi}$ in section 2.1.

Let P^A be the proportion of students meeting the standard z_A for time point A and P^B be the proportion for time point B. The two empirical curves on the left side of Figure 2 illustrate a change between two time points, whereas $\hat{\xi}^A$ and $\hat{\xi}^B$ on the right side of the figure are the results of mapping procedure.

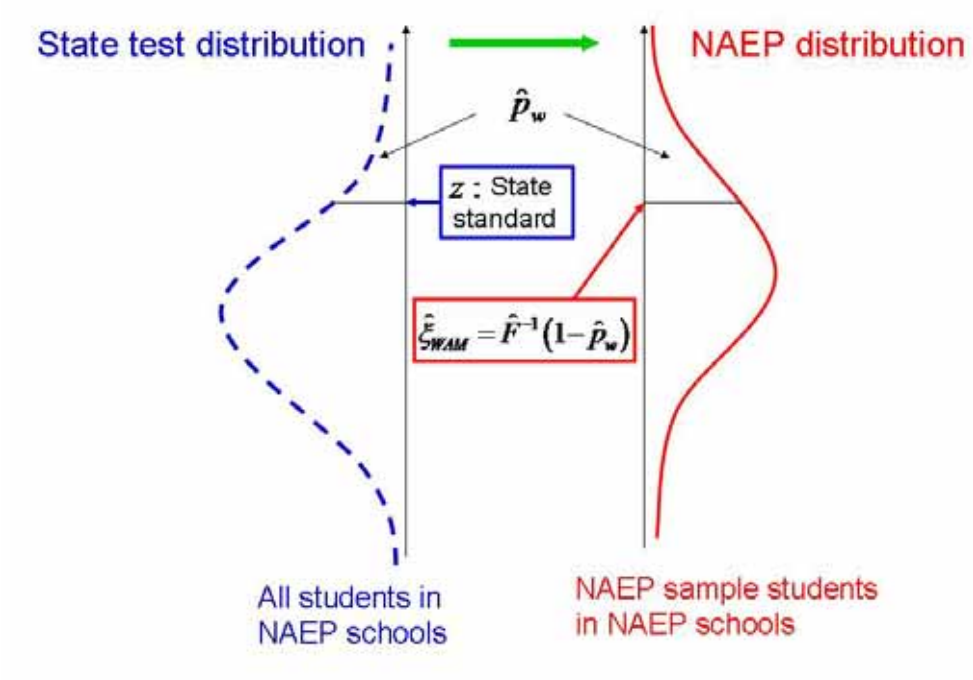

Figure 1. **The schematic of the mapping procedure.**

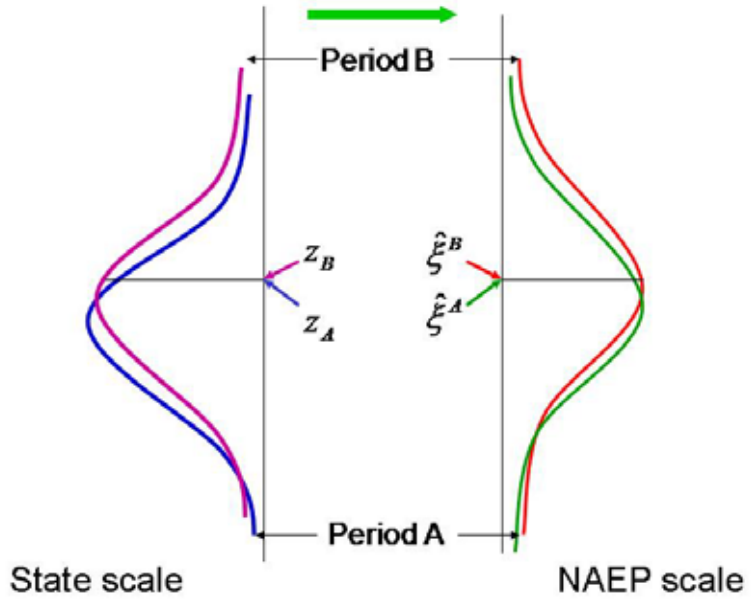

Figure 2. **The mapping procedure for a state test over two time periods.**

Let $P^B = P^A + \Delta P^S$, where ΔP^S is the change in the proportion of students meeting the standard in the state test. When $\Delta P^S > 0$, it means that a higher proportion of students met the standard at time point B. A higher proportion meeting the standard at the time point B could occur for one of two possible reasons: there is real progress in education or there is DPS in the testing results. If there is progress in education, it is assumed that the students should show a similar degree of progress in both the state test and corresponding NAEP assessment.

Some properties of the NAEP equivalents over time. Let F^A and F^B denote the estimated distributions on the NAEP scale for time point A and time point B. As given in (1), the NAEP equivalent for time point A, the image of P^A, is the $(1-P^A)$th quantile on F^A:

$$\xi^A = F^{-1,A}\left(1-P^A\right), \tag{2}$$

and the image of P^B on F^B is

$$\xi^B = F^{-1,B}\left(1-P^B\right) = F^{-1,B}\left(1-\left(P^A+\Delta P^S\right)\right). \tag{3}$$

Let P^α be the true proportion of students whose scores are greater than the point of ξ^A in the NAEP assessment at time point B, that is,

$$\xi^A = F^{-1,B}\left(1 - P^\alpha\right). \tag{4}$$

Because of the changes in performance over time, P^α is usually not equal to P^A. Thus $P^\alpha = P^A + \Delta P^N$, where ΔP^N is the changed proportion in NAEP above ξ^A at time point B.

First, assume $\Delta P^S = \Delta P^N$, i.e. for the time period students show the same change in achievement in both the state test and the corresponding NAEP assessment. It implies that $P^\alpha = P^A + \Delta P^S$. Because of (4) and

$$\xi^B = F^{-1,B}\left(1 - \left(P^A + \Delta P^S\right)\right) = F^{-1,B}\left(1 - P^\alpha\right), \tag{5}$$

thus $\xi^B = \xi^A$. This outcome indicates that when $\Delta P^S = \Delta P^N$, the NAEP equivalent is invariant over time. Accordingly, ξ^A can be viewed as being an *invariant equivalent*. Figure 2 illustrates how the mapping procedure for both the state test and the NAEP assessment performs for the time period in question. Using the NAEP scale as the benchmark for comparison, invariance of NAEP equivalents over time under the standard conditions is equivalent to the invariance of state proficiency standards over time.

Second, assume $\Delta P^S > \Delta P^N$, i.e. $P^A + \Delta P^S > P^A + \Delta P^N$, the proportion of students meeting the standard on the state test is higher than that on the NAEP assessment. Because

$$\xi^A = F^{-1,B}\left(1 - P^\alpha\right) = F^{-1,B}\left(1 - \left(P^A + \Delta P^N\right)\right) \tag{6}$$

and the monotone property of $F^{-1,B}(\bullet)$, it follows that

$$\xi^B = F^{-1,B}\left(1 - \left(P^A + \Delta P^S\right)\right) < F^{-1,B}\left(1 - \left(P^A + \Delta P^N\right)\right), \tag{7}$$

that is, $\xi^B < \xi^A$, which indicates that the NAEP equivalent at time point B is lower than ξ^A. It shows an occurrence of DPS, a deviation in proficiency standard. Figure 3 illustrates the empirical mapping procedure indicating that the state test performs differentially from the NAEP assessment over time.

Third, assume $\Delta P^S < \Delta P^N$, i.e. $P^A + \Delta P^S < P^A + \Delta P^N$, the proportion of students meeting the standard in the state test is lower than that in the NAEP assessment. Because of (6) and the monotone property of $F^{-1,B}(\bullet)$, it follows that

$$\xi^B = F^{-1,B}\left(1 - \left(P^A + \Delta P^S\right)\right) > F^{-1,B}\left(1 - \left(P^A + \Delta P^N\right)\right), \tag{8}$$

that is, $\xi^B > \xi^A$. This is a trivial case, though it shows an occurrence of DPS.

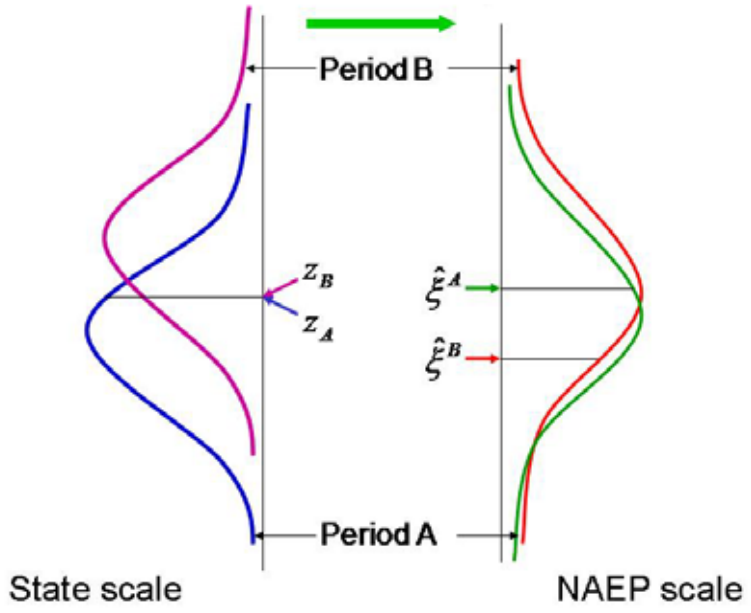

Figure 3. **The mapping procedure with score inflation in the state test for Period B.**

Test of the invariance of NAEP equivalents over time. In this study, the evaluation procedure employs both statistical significance tests and effect size criteria. For the statistical approach, the hypothesis serves as a check of the invariance of the NAEP equivalents over time under the standard conditions. The null hypothesis can be expressed as H_o: $\xi^B = \xi^A$. An equivalent hypothesis is whether the proportion of students passing ξ^B at time point B equals the proportion passing the invariant equivalent at time point B: $P^B = P^\alpha$. In this study, two significance tests are employed in this analysis. The first test uses a t-type statistic to check the difference of two proportions. The second statistic is the log-odds ratio (Haberman, 1978).

Let $n_{B.}$ be the sample size in consideration for time point B. Let $\hat{\xi}^B$ and $\hat{\xi}^A$ be the estimates of ξ^B and ξ^A, respectively. In Table 1, let n_{11} and n_{21} be the numbers of students whose scores are greater than $\hat{\xi}^B$ and $\hat{\xi}^A$, respectively, and n_{12} and n_{22} be the numbers of students who fail to meet the standards. Let $\hat{p}_w^B = n_{11}/n_{B.}$ be the estimate of P^B, and $\hat{p}_w^\alpha = n_{21}/n_{B.}$ be the estimate of P^α. Let $\hat{p} = n_{.1}/n$ and $\hat{q} = 1 - \hat{p}$. Define the Z_c statistic as

$$Z_c = \frac{\left|\hat{p}_w^{\alpha} - \hat{p}_w^{B}\right| - 1/n_{B.}}{\sqrt{2\,\hat{p}\hat{q}/n_{B.}}}. \tag{9}$$

The term, $1/n_{B.}$, in (9) is the Yates correction for continuity (Yates, 1934). The log-odds ratio is defined as

$$L = \log\left(\frac{n_{11}n_{22}}{n_{12}n_{21}}\right), \tag{10}$$

and an estimate of its standard error is

$$SE(L) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \tag{11}$$

Because the NAEP state data are collected by a two-stage sampling approach, the formulas for simple random sampling will underestimate the variances employed in the test statistics. The variance estimation for complex data usually uses replicate resampling approaches (Wolter, 1985). To simplify computations and count the effects of complex sampling, w the variances are estimated by multiplying a variance estimate by a design effect, which was introduced by Kish (1965) as a ratio of the variance of a statistic from complex samples over the variance of the statistic from simple random samples. Based on previous NAEP analyses, 2.5 is used as the approximate design effect for computation purposes. A .05 alpha level is then used in the analyses of the statistical tests.

Table 1

Number of NAEP Students Whose Scores Are Greater Than $\hat{\xi}^B$ and Passing $\hat{\xi}^A$ at Time Point B

	Proficiency standard		
	Pass	Fail	Total
$\hat{\xi}^B$	n_{11}	n_{12}	$n_{B.}$
$\hat{\xi}^A$	n_{21}	n_{22}	$n_{B.}$
Total	$n_{.1}$	$n_{.2}$	N

When the hypothesis is rejected, it shows a significant difference in the NAEP equivalents over time, which implies a significant DPS. It shows that the students in the state test have performed differently from how they would on the NAEP assessment. However, DPS cannot be considered equivalent to score inflation, because other possible factors could cause such differences, including differential performance gains and style of classroom instruction. Only when other potential factors can be dismissed can DPS be taken as an indication of score inflation.

For practical purposes, the effect size criterion is also used to evaluate the difference of two proportions drawn from independent samples, or the differences between a single proportion and any specified hypothetical value. The effect size for comparison of proportions is called the *H* index. To provide a better scale for looking at differences on which effect sizes for proportions are comparably detectable, Cohen (1988) applied the arcsine transformation to the proportions before calculating their difference. Let the arcsine transformation be $\varphi = 2\arcsin\sqrt{p}$. The *H* index for proportions is then defined as $H = |\varphi_1 - \varphi_2|$. To count as an intermediate effect size, the absolute value of the *H* index has to be at least 0.20 in the measuring differences of two proportions.

2.3 Evaluating the Test Results

After a significant DPS has been detected, it is also important to find the causes of this deviation in proficiency standards. To assess the test results, a committee of test experts and subject-matter specialists should be assembled. This process is analogous to the sort of review process that happens with a NAEP DIF analysis (Allen et al., 2001).

To judge the causes of the deviation in the NAEP equivalents over time, the whole process consists of two phases. The initial phase involves executing the relevant computations and statistical tests. The second phase involves assessing the results and determining the factors that could cause the deviation in proficiency standards over time. The expert committee will check the standard condition assumptions, review the results of the findings, discuss the possible causes for the differences, and draw conclusions. Only if all competing potential causes are eliminated can the results be attributed to score inflation.

3. Application to Empirical Data

3.1 Data

To detect the deviation in proficiency standard over time, two sets of data have been analyzed in this study: (a) the 2003 and 2005 NAEP mathematics and reading assessment samples for Grade 4 (G4) and Grade 8 (G8) students, and (b) the 2003 and 2005 state test samples of mathematics and reading for G4 and G8 students. Information on the proportions of students meeting state test standards for 2003 and 2005 was retrieved from the National Longitudinal School-Level State Assessment Score Database (NLSLSASD). This database contains the proportions of students, by school, meeting each of the state's standards for nearly all states, beginning as early as the academic year 1994. However, it does not contain scores for individual students. Typically, the NLSLSASD presents for each school the percent of students meeting or exceeding each achievement standard established by the state.[2]

3.2 Empirical Results

The mapping procedure described in section 2.1 was first completed. In Appendix B, Tables B1, B3, B5, and B7 display the estimates of the statewide proportion of proficient students, the estimated NAEP equivalents to the state standard, and the estimated standard error of the NAEP equivalent for the 2005 G4 and G8 reading and mathematics state tests, respectively. Tables B2, B4, B6, and B8 contain the same results for the 2003 G4 and G8 reading and mathematics state tests, respectively. For each state, each table also displays the number of schools in the NAEP sample and the number of schools employed in the mapping. This last quantity is simply the number of schools in the NAEP sample that could be matched to the schools with usable state test performance data. The notes under each of the tables list issues concerning data in this analysis.

In the G4 reading analysis, data from 21 states were used in the comparison, among 25 states having both 2005 and 2003 data. To align the state test and NAEP reading assessments, state data was dropped if the relevant state assessment was labeled "English/Language Arts" rather than "Reading." Furthermore, only those states that showed an increase in the proportion of students meeting their standards were discussed. The outcome shows two states having significant results in both statistical tests and effect size check. These 2 states are listed in Table 2 (States 1 and 2).

Table 2

Results of Statistical Testing and H Index Checking With Significance for the Grade 4 Reading and Mathematics

State	2005: estimate of proportion passing $\hat{\xi}^B$, $\hat{\bar{p}}_w^B$	2005: estimated NAEP equivalent, $\hat{\xi}^B$	2005: estimate of proportion passing $\hat{\xi}^A$, $\hat{\bar{p}}_w^\alpha$	2003: estimated NAEP equivalent, $\hat{\xi}^A$	z_c statistic	Log-odds ratio	*H* index
Grade 4 reading:							
1	0.71	202	0.60	212	6.61	0.21	0.23
2	0.80	197	0.67	210	6.89	0.29	0.30
Grade 8 reading:							
3	0.63	244	0.52	256	5.15	0.19	0.22
4	0.82	235	0.73	247	5.24	0.23	0.22
5	0.72	245	0.63	256	5.56	0.18	0.19
6	0.30	276	0.19	285	6.02	0.27	0.26
7	0.57	254	0.43	267	6.48	0.25	0.28
Grade 4 math:							
8	0.85	218	0.76	226	5.72	0.25	0.23
9	0.80	224	0.65	234	6.79	0.33	0.34
10	0.91	207	0.78	217	8.50	0.45	0.37
Grade 8 math:							
11	0.61	269	0.52	278	4.35	0.18	0.20
12	0.53	276	0.44	286	4.51	0.17	0.20
13	0.74	258	0.64	268	4.68	0.20	0.22
14	0.70	266	0.53	280	8.24	0.32	0.35
15	0.65	277	0.44	293	8.82	0.37	0.42

Note that the names of all the states listed in Table 2 are unspecified, because the possible causes for the deviation in proficiency standards have not yet been investigated. For example, the State 1 test shows a large increase in the proportions of students meeting its standards. In the 2005 NAEP sample for State 1, the proportion of the students who passed ξ^B is .71 and the proportion of those who passed ξ^A is about .60. The images of $\overline{p}_w^A$ on $\hat{F}^A$ (.60) and $\overline{p}_w^B$ on $\hat{F}^B$ (.71) show significant variation in the NAEP scale over time. This indicates the presence of a significant DPS, or a deviation in state proficiency standards. In the G8 reading analysis, data from 28 states were used in the comparison, among 30 states having both 2005 and 2003 data. The outcome shows five states (States 3-7) having significant results in both statistical testing and the effect size check. Table 2 displays the results for these five states.

In the G4 mathematics analysis, among the 25 states having both 2005 and 2003 data, the data from 24 states are used in the comparison. After the first phase of the analysis, three states (States 8-10) listed in Table 2 show significant difference in the NAEP equivalents in the statistical testing and effect size checks. Among them, the State 8 test shows a substantial increase in the proportions of students meeting its standards. There were 74 % and 85% of the students passing its standard in 2003 and 2005, respectively. In the 2005 NAEP sample for State 8, the proportion of the students who passed its ξ^A is .76. The tests show that the variation of the images of $\overline{p}_w^A$ on $\hat{F}^A$ (.76) and $\overline{p}_w^B$ on $\hat{F}^B$ (.85) is significant. It implies that the G4 State 8 mathematics test shows a significant DPS, a deviation in state proficiency standards. To confirm the cause for these changes in achievement level percentages, further investigation must be conducted and final approval must be acquired from an expert committee in a second phase analysis. In the G8 mathematics analysis, data from 25 states were used in the comparison, among 32 states having both 2005 and 2003 data. The outcome shows five states (States 11-15) having significant results in both statistical testing and effect size check.

4. An Application: Detecting Score Inflation for State Tests

An important application of this approach is to detect score inflation for state tests. If other factors causing DPS can be excluded, a significant DPS indicates score inflation. So DPS is a necessary condition for the demonstration of score inflation.

In recent years, score inflation became a concern to many educators, because it has compromised efforts to improve education and accountability in assessments (Bromley, Crow, & Gibson, 1973; Hambleton et al., 1995; Rosovsky & Hartley, 2002; Shepard, 1988). Score inflation could be tied to a variety of situations. Clearly, for nonlinked or poorly equated tests, the lack of adequate equating could result in what might be considered to be grade inflation. But even if the scale of a test is well linked or equated, score inflation could still be present. A typical situation occurs when classroom instruction is test-driven and/or students are focused on learning the specific content in the questions on a standardized test. Because students at different achievement levels are part of this study of the content of questions, the resulting scores will not necessarily indicate the real academic level of individual students. In particular, students at a lower proficiency level often achieve test scores that are higher than their relative aptitude in such environments (Haladyna, Nolan, & Haas, 1991; Madaus, 1988a; Phelps, 2005). Such situations would result in the failure of the assessments to adequately measure student levels of achievement; even efforts to align tests closely with curricular standards would be insufficient to guard against this sort of score inflation (Koretz, 2005).

The principle in testing for score inflation is to check whether the score improvements on state tests can be corroborated by score improvements on NAEP. The stability of NAEP scales is the basis of such comparisons. If a DPS has been detected, a panel is then asked to determine if the cause of the deviation is likely due to score inflation.

Of the two cases of DPS discussed in section 2.2, only one gives an indicator of score inflation. When $\Delta P^S > \Delta P^N$, it implies that the proportion of students meeting the standard on the state test is higher than that on the NAEP assessment. The NAEP equivalent at time point B is lower than ξ^A; $\xi^B < \xi^A$. This case of significant DPS provides a scenario for possible score inflation. For another case, when $\Delta P^S < \Delta P^N$, it implies $\xi^B > \xi^A$. Such a significant DPS is not an indicator for score inflation. It may be due to failure to satisfy standard conditions or a change of testing conditions.

To formally claim score inflation, the causes for DPS must be evaluated by an expert committee, and the potential factors other than score inflation must be discussed. In addition, it is possible that the changes in the NAEP equivalents over time may be caused by a combination of factors: they could be partly due to modification of the item formats and test structures and partly

due to score inflation. Resolving this situation and drawing conclusions will necessitate the collection of additional data in further studies.

Although analysis of the 2005 and 2003 G4 and G8 reading and mathematics data in Table 2 has demonstrated significant DPS, one is unable to claim specific causes of DPS, including possible score inflation, because these results have not been reviewed by an expert committee.

5. Conclusions

This paper has developed an approach for testing the invariance of state proficiency standards over time for state tests or other analogous assessments. The approach is based on the methodology originally developed for making useful comparisons among state standards; the NAEP scale was used as the benchmark in both the original and the current development.

The approach arises from the need to deal with a practical testing issue (Thissen, 2007). It is well known that over time, factors such as score inflation, scale drift, differential performance gains, test instrument structure changes, content modification, and style of classroom instruction could cause a deviation in test scores. Apparently, this concept is broader than a deviation in proficiency standards. Table 3 shows an example of a deviation in test scores. It lists the changes in the fourth grade reading score in Kentucky's KIRIS[3] and its corresponding NAEP score (Koretz, 2007).

It is evident that the NAEP and KIRIS show different patterns of change, but a direct comparison does not make sense because they are measured on different scales. Although a standardized transformation could align the two scales, the test statistic based on the standardized transformation is rather complex. Instead, this issue can be detected by the methodology proposed in this study.

Table 3

Changes in Fourth-Grade Reading Proficiency in Kentucky's KIRIS and NAEP (1992–1994)

	Raw change	Standardized change
KIRIS	18.8	0.76
NAEP	-1.0	-0.03

The entire process comprises detecting DPS over time, verifying that standard test conditions were met, and evaluating the causes of changes by an expert committee. Under standard conditions, substantial difference in NAEP equivalents over time is an indicator of possible score inflation. In general, for any test that is statistically significant as reported by this method, a committee of test experts and subject-matter specialists is needed to review the results and determine whether score inflation can be considered to be the cause. This can be determined only after the other factors related to the changes in test conditions, such as content modification and changes in test instruments have been discounted. For the Kentucky example, in order to confirm whether score inflation is the cause for the KIRIS variance in test score, the entire statistical process suggested earlier should be executed.

As mentioned in section 2.3, it is possible that the differentiation of the NAEP equivalents over time may be caused partly by changes in test conditions and partly by score inflation. The existence of a combination of causes will add difficulty to the investigation. With only limited information available, any inferences concerning score inflation in this scenario should be made with due caution.

References

Allen, N., Donoghue, J., & Schoeps, T. (2001). *The NAEP 1998 technical report* (NCES 2001-509). Washington DC: National Center for Education Statistics.

American Federation of Teachers (2006). *Smart testing: Let's get it right*. Unpublished reviews, available from http://www.aft.org/pubs-reports/downloads/teachers/Testingbrief.pdf

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Arenson, K. W. (2004, April 18). Is it grade inflation, or are students just smarter? *New York Times*, p. WK2.,

Braun, H. I. & Qian, J. (2007a). An enhanced method for mapping state standards onto the NAEP scale. In N. J. Dorans, M. Pommerich, & P. W. Holland, (Eds.), *Linking and aligning scores and scales* (pp. 313–338). New York: Springer-Verlag.

Braun, H. I., & Qian J. (2007b). *Mapping 2005 state proficiency standards onto the NAEP scales* (NCES Research and Development Report No. NCES 2007–482). Washington DC: National Center for Education Statistics.

Bromley, D. G., Crow, H. L., & Gibson, M. S. (1973). Grade inflation: Trends, causes, and implications. *Phi Delta Kappan, 59*(10), 694–697.

Cannell, J. J. (1987). *Nationally normed elementary achievement testing in America's public schools. How all fifty states are above the national average*. Daniels, WV: Friends for Education.

Cochran, W.G. (1977). *Sampling techniques* (3rd ed.). New York: John Wiley & Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Feuer, M. J., Holland, P., Green, B. F., Bertenthal, M. W., & Hemphill, F. (Eds.). (1998). *Uncommon measures: Equivalence and linkage among educational tests*. Washington, DC: National Academy of Science.

Grissmer, D., Flanagan, A., Kawata, J., & Williamson, S. (2000). *Improving student achievement: What state NAEP test scores tell us*. (Rand Corporation Rep. No. MR-924-EDU). Santa Monica, CA: Rand Corporation.

Haberman, S. J. (1978). *Analysis of qualitative data: Vol. 1, Introductory topics*. New York: Academic Press.

Haladyna, T., Nolen, S., & Haas, N. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20*(5), 2-7.

Hambleton, R. K., Jaeger, R. M., Koretz, D., Linn, R. L., Millman, J., & Phillips, S. E. (1995). *Review of the measurement quality of the Kentucky Instructional Results Information System, 1991-1994.* Frankfort, KY: Office of Education Accountability, Kentucky General Assembly.

Haney, W. (2002). Ensuring failure: How a state's achievement test may be designed to do just that. *Education Week*, *56*, 58.

Jones, L., & Olkin, I. (2004). *The nation's report card: Evolution and perspectives*. Bloomington, Indiana: Phi Delta Kappa International.

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons.

Klein, S. P., Hamilton, L. S., McCaffrey, D. F., & Stecher, B. M. (2000). What do test scores in Texas tell us? *Education Policy Analysis Archives*, *8*, 49.

Koretz, D. M. (1988). Arriving in Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? *American Educator*, *12*(2), 8-15, 46-52.

Koretz, D. M. (2005). Alignment, high stakes, and the inflation of test scores. *Yearbook of the National Society for the Study of Education 104*(2), 99–118.

Koretz, D. M. (2007). Using aggregate-level linkages for estimation and validation: Comments on Thissen and Braun & Qian. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 339–353). New York: Springer-Verlag.

Koretz, D. M., Bertenthal, M. W., & Green, B. F. (Eds.). (1999). *Embedding questions: The pursuit of a common measure in uncommon tests.* Washington, DC: National Academy of Sciences.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing State and district results to national norms: The validity of the claims that "Everyone is above average." *Educational Measurement: Issues and Practice, 9*(3), 5–14.

Madaus, G. F. (1988a). The distortion of teaching and testing: High-stakes testing and instruction. *Peabody Journal of Education*, *65*(3), 29–46.

Madaus, G. F. (1988b). The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical issues in curriculum* (pp. 83–121). Chicago: University of Chicago Press.

McLaughlin, D., & Bandeira de Mello, V. (2003, June). *Comparing state reading and math performance standards using NAEP*. Paper presented at the National Conference on Large-Scale Assessment, San Antonio, TX.

Neill, M., & the Staff of FairTest. (1997). *Testing our children: A report card on state assessment systems*. Cambridge, MA: National Center for Fair & Open Testing.

Phelps, R. P. (2005). The source of Lake Wobegon. *Third Education Group Review*, *1*, 2.

Potter, W. P. (1979). Grade inflation: Unmasking the scourge of the seventies. *College and University, 55*(1), 19–26.

Qian, J., Kaplan, E., Johnson, E., Krenzke, T., & Rust, K. (2001). State weighting procedures and variance estimation. In N. Allen, J. Donoghue, & T. Schoeps (Eds.), *The NAEP 1998 technical report* (pp. 193–225). Washington, DC: National Center for Education Statistics.

Rosovsky, H. & Hartley, M. (2002). *Evaluation and the Academy: Are we doing the right thing? Grade inflation and letters of recommendation*. Cambridge, MA: American Academy of Arts and Sciences.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.

Shepard, L. A. (1988, April). *The harm of measurement-driven instruction.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Skinner, C., Holt, D., & Smith, T. (1989). *Analysis of complex surveys.* New York: John Wiley & Sons.

Smith, M. L. (1991). Meanings of test preparation. *American Educational Research Journal*, *28*(3), 521–542.

Thissen, D. (2007). Linking assessments based on aggregate reporting: Background and issues. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and aligning scores and scales* (pp. 287–312). New York: Springer-Verlag.

Wolter, K. (1985). *Introduction to variance estimation*. New York: Springer-Verlag.

Yates, F. (1934). Contingency table involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society* (Supplement) *1*, 217–235.

Notes

[1]The National Longitudinal School-Level State Assessment Score Database (NLSLSASD; www.schooldata.org) is constructed and maintained by the American Institutes for Research (AIR) for the National Center for Education Statistics (NCES). Its purpose is to collect and validate data from state testing programs across the country. It contains assessment data for approximately 80,000 public schools in the United States and is updated annually.

[2] For almost all states, some schools in the NAEP school sample were either missing from the NLSLSASD or the required datum was not listed. In those cases, the number of schools available for estimation was smaller than the number of schools in the NAEP school sample. For each subject and grade combination, there were four to five jurisdictions in which the proportion of NAEP sample schools employed in the estimation was less than 0.9.

[3] The Kentucky Instructional Results Information System (KIRIS). It was replaced by the Commonwealth Accountability and Testing System (CATS) in the 1998-99 school year.

[4] Students with disabilities and English language learners who cannot be assessed, even with the accommodations that NAEP provides, are not considered nonrespondents, but are excluded from the population of inference. Their performance is not included in estimates of the NAEP score distributions.

[5] Note that this calculation was carried out only for the subset of NAEP sample schools with complete data. School and student weights were not adjusted for schools lost from the NAEP school sample due to nonresponse.

Appendix A

NAEP Sample Design, School Weights, and Target Estimation

NAEP Sample Design and School Weights

State NAEP samples are obtained through a two-stage probability sampling design. The first stage constitutes a probability sample of schools containing the relevant grade. The second stage involves the selection of a random sample of students within each school. To account for the unequal probabilities of selection and to allow for adjustments for nonresponse, each school and each student was assigned a separate sampling weight.[4] If these weights are not employed in the computation of the statistics of interest, the resulting estimates can be biased. With this caution in mind, appropriate weights were applied in the estimation of the proportion of students in the state above the standard. In general, the school weight equals the inverse of the approximate school selection probability, and the student weight is inversely proportional to the product of the school selection probability and the student selection probability. A more detailed description of school weights can be found in Braun & Qian (2007a).

Because school weights are not retained in the NAEP database, for this study the school weights were computed in two steps. First, the sum of the student design weights for each school was calculated and then this sum was divided by the number of grade-eligible students.[5] Details of the creation of school design weights for NAEP can be found in *NAEP 1998 Technical Report* (Qian, Kaplan, Johnson, Krenzke, & Rust, 2001, Chap. 11).

The Ratio Estimator for the Target Proportion

Let P_k be the proportion of students achieving the standard at school *k* and w_k be the corresponding school weight. The total number of students meeting the standard is $\sum_{l=1}^{N} P_l \cdot M_l$, where *N* is the total number of public schools in the state containing the relevant grade and M_l is the number of students who were grade-eligible at school *l*, (including all students with disabilities and English language learners). The statewide target proportion of students meeting the standard is approximately

$$P = \frac{\sum_{l=1}^{N} P_l \cdot M_l}{\sum_{l=1}^{N} M_l}.$$

Using Horvitz–Thompson estimators (Cochran, 1977), the numerator and denominator of *P* are estimated separately from the state's NAEP school sample. For example, $\sum_{l=1}^{n} w_l M_l$ estimates the total number of eligible students in the state, and $\sum_{l=1}^{n} w_l \left(P_l \cdot M_l\right)$ estimates the total number of students meeting the standard. The target proportion, *P*, of students meeting the standard can be estimated by a ratio estimator:

$$\overline{p}_w = \frac{\sum_{l=1}^{n} w_l \left(P_l \cdot M_l\right)}{\sum_{l=1}^{n} w_l M_l}.$$

Variance Estimation

When survey variables are observed without error from every respondent to a stratified and clustered sample as NAEP is, the usual complex-sample variance estimators quantify the uncertainty associated with sample statistics (Skinner, Holt, & Smith, 1989). The fact that a specific NAEP score is not assigned to individual students participating in the NAEP assessments (even those who responded to the cognitive items), requires additional statistical analyses to properly quantify the uncertainty associated with inferences about score distributions (Allen et al., 2001; Wolter, 1985).

The total variance of the estimate of the NAEP equivalent to a state standard consists of two components: (a) the error due to sampling schools and students and (b) the error of measurement that reflects the uncertainty in an assessed student's performance. The sampling error is estimated by applying the jackknife replicate resampling (JRR) approach to the mapping procedure. The estimation involves the corresponding schools on the state data and on the NAEP data. The measurement error due to unobservability is estimated by utilizing the variability among the five sets of plausible values generated for each assessed student (Rubin, 1987).

Appendix B

Results of Mapping State Standards

Table B1

Results of Mapping State Standards to the Grade 4 NAEP Reading Scale: 2005

State	State name	# of schools in NAEP sample	# of schools in mapping	Estimated proportion meeting state proficiency standard, $\hat{\bar{p}}_w^B$	Estimated NAEP equivalent to state standard, $\hat{\xi}^B$	Estimated standard error of NAEP equivalent, $se\left(\hat{\xi}^B\right)$
AK	Alaska[a]	157	97	0.79	182	2.6
AR	Arkansas	151	144	0.53	217	1.2
CA	California	445	421	0.48	210	0.9
CO	Colorado	147	135	0.86	186	1.6
CT	Connecticut	132	132	0.66	212	1.0
FL	Florida	169	159	0.71	202	1.0
GA	Georgia[a]	176	156	0.87	175	2.2
HI	Hawaii	132	131	0.56	205	1.1
IA	Iowa	130	125	0.77	197	1.2
ID	Idaho	157	148	0.87	185	2.9
IN	Indiana	138	138	0.72	199	1.1
KY	Kentucky	149	148	0.67	206	1.6
LA	Louisiana	136	134	0.65	198	2.0
MA	Massachusetts	202	199	0.48	234	0.8
MD	Maryland	125	123	0.82	187	1.4
MS	Mississippi	127	116	0.88	161	2.0
MT	Montana[a]	241	194	0.81	197	1.5
NC	North Carolina	175	168	0.82	183	1.6
ND	North Dakota[a]	261	194	0.76	204	0.8
NJ	New Jersey	135	134	0.81	191	1.6
NM	New Mexico[a]	161	135	0.50	208	1.2
NV	Nevada	120	113	0.48	212	1.4
NY	New York	190	186	0.71	207	1.5
OH	Ohio	201	198	0.77	199	1.9
OK	Oklahoma	176	175	0.82	182	1.8
SC	South Carolina	119	118	0.35	228	1.3
TN	Tennessee	139	137	0.88	170	2.3
TX	Texas	383	376	0.81	190	1.0
WA	Washington	136	133	0.80	197	1.6
WI	Wisconsin	169	169	0.83	189	1.8
WV	West Virginia	195	190	0.80	186	1.3
WY	Wyoming[a]	170	146	0.47	228	0.7

Note. NAEP reading cut scores at Grade 4 are 208 for *Basic* and 238 for *Proficient.* The following states' Grade 4 reading test data were not used in the analysis or received special treatment: ME and MI—results deleted due to discrepancies between state assessment data and the state document; CA and LA—reading data not available for the state assessment, so English Language Arts (ELA) data used; MA—reading data not available for state assessment, so "Language" data variable used; AZ, DC, DE, IL, KS, MN, MO, OR, PA, and VA—neither reading nor ELA data available in the state data file; AL, NH, RI, SD, UT, and VT—state assessment data not available; NE—state results are based on assessments developed by each local education agency. SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Reading Assessment, and National Longitudinal School-Level State Assessment Score Database (NLSLSASD).

[a] The proportion of NAEP sample schools employed in the estimation was less than 0.9.

Table B2

Results of Mapping State Standards to the Grade 4 NAEP Reading Scale: 2003

State	State name	# of schools in mapping	Estimated proportion meeting state proficiency standard, $\hat{\bar{p}}_w^B$	Estimated NAEP equivalent to state standard, $\hat{\xi}^B$	Estimated standard error of NAEP equivalent, $se\left(\hat{\xi}^B\right)$
AK	Alaska	103	0.73	193	2.6
AR	Arkansas	115	0.62	206	1.7
CA	California	216	0.38	219	1.3
CO	Colorado	115	0.87	184	2.1
CT	Connecticut	108	0.68	215	1.8
DC	District of Columbia	103	0.47	192	0.8
FL	Florida	104	0.58	212	1.3
GA	Georgia	147	0.80	183	1.6
IA	Iowa	129	0.77	201	1.8
ID	Idaho	114	0.75	197	1.6
KY	Kentucky	121	0.62	211	1.6
LA	Louisiana	109	0.59	198	2.0
MA	Massachusetts	161	0.54	226	1.4
ME	Maine	145	0.50	226	1.1
MI	Michigan	133	0.74	197	1.8
MS	Mississippi	107	0.87	165	1.7
MT	Montana	141	0.77	199	2.0
NC	North Carolina	147	0.81	191	1.4
ND	North Dakota	176	0.75	201	1.0
NJ	New Jersey	109	0.78	198	1.2
NV	Nevada	106	0.49	211	1.5
NY	New York	145	0.64	211	1.6
OH	Ohio	163	0.69	207	2.2
SC	South Carolina	101	0.31	234	1.7
TX	Texas	194	0.85	177	1.7
WA	Washington	95	0.65	210	1.3
WI	Wisconsin	127	0.82	190	1.2
WY	Wyoming	145	0.44	230	1.0

Note. NAEP reading cut scores at Grade 4 are 208 for *Basic* and 238 for *Proficient.* (Median SE of the NAEP equivalent = 1.6.) CA and LA—reading data not available for the state assessment, so English Language Arts (ELA) data used. SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Reading Assessment, and National Longitudinal School-Level State Assessment Score Database (NLSLSASD).

Table B3

Results of Mapping State Standards to the Grade 8 NAEP Reading Scale: 2005

State	State name	# of schools in NAEP sample	# of schools in mapping	Estimated proportion meeting state proficiency standard, $\hat{\bar{p}}_w^B$	Estimated NAEP equivalent to state standard, $\hat{\xi}^B$	Estimated standard error of NAEP equivalent, $se\left(\hat{\xi}^B\right)$
AK	Alaska[a]	102	54	0.82	230	1.2
AR	Arkansas[a]	125	112	0.57	254	1.2
AZ	Arizona	132	125	0.63	244	1.3
CA	California	374	356	0.39	262	0.8
CO	Colorado	120	108	0.86	229	2.1
CT	Connecticut	106	102	0.77	242	1.7
DC	District of Columbia[a]	42	28	0.44	244	0.9
DE	Delaware[a]	43	37	0.80	242	0.9
FL	Florida	161	155	0.44	265	1.5
GA	Georgia	124	116	0.83	224	2.2
HI	Hawaii	67	64	0.37	262	1.4
IA	Iowa	111	109	0.72	250	1.0
ID	Idaho	101	93	0.82	235	2.5
IL	Illinois	190	187	0.72	245	1.2
IN	Indiana	107	105	0.66	249	1.5
KS	Kansas	117	114	0.78	242	1.4
LA	Louisiana	112	110	0.54	251	1.4
MD	Maryland	107	105	0.68	245	1.7
MS	Mississippi	115	104	0.58	247	1.4
NC	North Carolina	139	132	0.88	217	1.5
ND	North Dakota[a]	182	134	0.72	255	0.9
NJ	New Jersey	111	110	0.74	250	1.3
NM	New Mexico[a]	106	86	0.52	251	1.2
NY	New York	182	173	0.49	268	1.1
OH	Ohio	142	135	0.80	241	1.5
OK	Oklahoma	147	142	0.71	244	1.9
OR	Oregon	119	116	0.64	254	1.3
PA	Pennsylvania	110	104	0.64	258	1.7
SC	South Carolina	108	104	0.30	276	1.3
TN	Tennessee	112	111	0.87	222	1.5
TX	Texas	278	270	0.83	225	1.0
WI	Wisconsin	118	117	0.86	229	2.1
WV	West Virginia	110	107	0.80	228	1.7
WY	Wyoming	78	77	0.39	278	1.2

Note. NAEP reading cut scores at Grade 8 are 243 for *Basic* and 281 for *Proficient.* The following states' Grade 8 reading test data were not used in the analysis or received special treatment: ME, MT, and VA—discrepancies exist between the state assessment data and the state document; CA and LA—reading data not available for state assessment, so ELA data used; KY, MA, MI, MN, MO, NV, and WA—neither reading nor ELA data available in the state data file; AL, NH, RI, SD, UT, and VT—state assessment data not available; NE—state results are based on assessments developed by each local education agency. SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Reading Assessment, and National Longitudinal School-Level State Assessment Score Database (NLSLSASD).

[a]The proportion of NAEP sample schools employed in the estimation was less than 0.9.

Table B4

Results of Mapping State Standards to the Grade 8 NAEP Reading Scale: 2003

State	State name	# of schools in mapping	Estimated proportion meeting state proficiency standard, $\hat{\bar{p}}_w^B$	Estimated NAEP equivalent to state standard, $\hat{\xi}^B$	Estimated standard error of NAEP equivalent, $se\left(\hat{\xi}^B\right)$
AK	Alaska	51	0.71	241	1.7
AR	Arkansas	99	0.44	267	1.8
AZ	Arizona	105	0.54	256	1.5
CA	California	180	0.32	271	1.2
CO	Colorado	104	0.88	229	1.9
CT	Connecticut	102	0.79	239	2.2
DC	District of Columbia	26	0.45	244	1.0
DE	Delaware	32	0.70	249	0.9
FL	Florida	96	0.47	263	1.6
GA	Georgia	113	0.81	230	2.1
HI	Hawaii	53	0.39	264	1.0
IA	Iowa	115	0.70	253	0.8
ID	Idaho	85	0.73	247	1.5
IL	Illinois	169	0.65	256	1.3
IN	Indiana	99	0.63	257	1.1
KS	Kansas	118	0.69	253	1.3
LA	Louisiana	94	0.52	253	1.5
MD	Maryland	95	0.62	252	1.7
ME	Maine	106	0.45	274	1.3
MS	Mississippi	102	0.55	250	1.3
MT	Montana	100	0.72	253	1.1
NC	North Carolina	129	0.86	226	1.6
ND	North Dakota	31	0.69	255	1.2
NJ	New Jersey	107	0.74	249	1.6
NY	New York	141	0.47	272	1.3
OK	Oklahoma	123	0.78	238	1.8
OR	Oregon	105	0.59	258	1.0
PA	Pennsylvania	100	0.63	256	1.5
SC	South Carolina	92	0.21	285	1.5
TX	Texas	142	0.88	221	1.7
WI	Wisconsin	103	0.84	232	2.9
WY	Wyoming	74	0.39	277	0.9

Note. NAEP reading cut scores at Grade 8 are 243 for *Basic* and 281 for *Proficient.* (Median SE of the NAEP equivalent = 1.5.) The following states' Grade 8 reading test data were not used in the analysis or received special treatment: VA—results deleted due to discrepancies between state assessment data and the state document; CA and LA—reading data not available for the state assessment, so English Language Arts (ELA) data used. SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Reading Assessment, and National Longitudinal School-Level State Assessment Score Database (NLSLSASD).

Table B5

Results of Mapping State Standards to the Grade 4 NAEP Mathematics Scale: 2005

State	State name	# of schools in NAEP sample	# of schools in mapping	Estimated proportion meeting state proficiency standard, $\hat{\tilde{P}}_w^B$	Estimated NAEP equivalent to state standard, $\hat{\xi}^B$	Estimated standard error of NAEP equivalent, $se\left(\hat{\xi}^B\right)$
AK	Alaska[a]	153	108	0.71	222	1.4
AR	Arkansas	151	144	0.53	236	1.0
CA	California	446	421	0.51	231	0.7
CO	Colorado	146	135	0.90	201	1.7
CT	Connecticut	132	132	0.78	221	1.0
FL	Florida	169	159	0.63	230	0.8
GA	Georgia[a]	176	156	0.75	215	1.4
HI	Hawaii	132	131	0.30	247	1.2
IA	Iowa	130	124	0.80	219	1.1
ID	Idaho	158	148	0.91	207	1.9
IN	Indiana	138	138	0.72	225	1.1
KS	Kansas	139	134	0.85	218	1.4
LA	Louisiana	136	134	0.63	223	1.0
MA	Massachusetts	202	200	0.39	255	1.0
MD	Maryland	125	124	0.78	215	1.1
MI	Michigan	141	131	0.73	222	1.7
MO	Missouri	159	158	0.41	242	1.2
MS	Mississippi	127	117	0.79	206	1.3
NC	North Carolina	175	168	0.91	203	1.2
ND	North Dakota[a]	261	194	0.80	224	0.8
NJ	New Jersey	135	134	0.81	221	1.3
NM	New Mexico[a]	162	135	0.39	233	1.3
NV	Nevada	118	112	0.52	230	0.9
NY	New York	190	186	0.87	207	1.5
OH	Ohio	201	199	0.65	233	1.3
OK	Oklahoma	177	175	0.74	218	0.9
SC	South Carolina	119	118	0.39	246	1.2
TN	Tennessee	139	137	0.87	200	1.6
TX	Texas	382	376	0.82	219	1.0
WA	Washington	136	133	0.60	236	1.1
WI	Wisconsin	169	169	0.74	225	1.4
WV	West Virginia	195	190	0.75	215	1.1
WY	Wyoming[a]	164	146	0.39	251	0.7

Note. NAEP mathematics cut scores at Grade 4 are 214 for *Basic* and 249 for *Proficient.* The following states' Grade 4 mathematics test data were not used in the analysis or received special treatment: ME and MT—discrepancies exist between the state assessment data and the state document; AZ, DC, DE, IL, KY, MN, OR, PA, and VA—data not available in the file; AL, NH, RI, SD, UT, and VT—state assessment data not available; NE—state results are based on assessments developed by each local education agency. SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Mathematics Assessment, and National Longitudinal School-Level State Assessment Score Database (NLSLSASD).

[a] The proportion of NAEP sample schools employed in the estimation was less than 0.9.

Table B6

Results of Mapping State Standards to the Grade 4 NAEP Mathematics Scale: 2003

State	State name	# of schools in mapping	Estimated proportion meeting state proficiency standard, $\hat{\bar{P}}_w^B$	Estimated NAEP equivalent to state standard, $\hat{\xi}^B$	Estimated standard error of NAEP equivalent, $se\left(\hat{\xi}^B\right)$
AK	Alaska	110	0.67	223	1.3
AR	Arkansas	115	0.60	223	0.9
CA	California	216	0.45	231	1.1
CT	Connecticut	108	0.80	217	1.1
DC	District of Columbia	103	0.54	201	0.7
FL	Florida	103	0.56	231	1.3
GA	Georgia	147	0.74	212	1.1
IA	Iowa	130	0.77	220	1.1
ID	Idaho	114	0.77	217	0.9
KS	Kansas	130	0.74	226	1.1
LA	Louisiana	109	0.58	221	1.1
MA	Massachusetts	161	0.38	251	1.1
ME	Maine	145	0.29	252	0.8
MI	Michigan	133	0.64	226	1.2
MO	Missouri	126	0.37	244	1.0
MS	Mississippi	107	0.74	205	1.3
MT	Montana	142	0.75	220	0.9
NC	North Carolina	151	0.92	203	1.0
ND	North Dakota	176	0.59	234	0.7
NJ	New Jersey	109	0.68	227	1.4
NV	Nevada	106	0.51	228	1.0
NY	New York	145	0.79	213	1.1
OH	Ohio	163	0.59	232	1.0
SC	South Carolina	101	0.33	248	0.9
TX	Texas	194	0.88	207	1.5
WA	Washington	96	0.54	236	1.2
WI	Wisconsin	127	0.70	223	1.1
WY	Wyoming	145	0.36	250	0.6

Note. NAEP mathematics cut scores at Grade 4 are 214 for *Basic* and 249 for *Proficient.* (Median SE of the NAEP equivalent = 1.1.) SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Mathematics Assessment, and National Longitudinal School-Level State Assessment Score Database (NLSLSASD).

Table B7

Results of Mapping State Standards to the Grade 8 NAEP Mathematics Scale: 2005

State	State name	# of schools in NAEP sample	# of schools in mapping	Estimated proportion meeting state proficiency standard, $\hat{\bar{P}}_w^B$	Estimated NAEP equivalent to state standard, $\hat{\xi}^B$	Estimated standard error of NAEP equivalent, $se\left(\hat{\xi}^B\right)$
AK	Alaska[a]	101	59	0.65	268	0.9
AR	Arkansas[a]	125	112	0.34	288	1.0
AZ	Arizona	131	125	0.61	265	1.1
CO	Colorado[a]	121	108	0.74	258	1.6
CT	Connecticut	106	102	0.76	257	2.3
DC	District of Columbia[a]	42	28	0.40	252	1.4
DE	Delaware[a]	43	37	0.56	275	1.0
FL	Florida	162	155	0.58	269	1.3
GA	Georgia	124	116	0.69	255	1.2
HI	Hawaii	67	64	0.20	296	1.2
IA	Iowa	111	109	0.76	262	1.1
ID	Idaho	103	93	0.70	266	1.7
IL	Illinois	190	187	0.54	276	1.5
IN	Indiana	107	105	0.70	266	1.5
KY	Kentucky	117	115	0.37	285	1.4
LA	Louisiana	112	110	0.56	264	1.6
MA	Massachusetts	131	128	0.42	301	1.3
MD	Maryland	107	105	0.53	276	1.7
MI	Michigan	116	111	0.61	269	1.9
MO	Missouri	131	129	0.15	311	1.4
MS	Mississippi	115	104	0.53	262	1.5
NC	North Carolina	140	133	0.84	247	1.2
ND	North Dakota[a]	184	135	0.65	277	1.1
NJ	New Jersey	111	110	0.64	273	1.4
NM	New Mexico[a]	106	86	0.24	287	1.8
NY	New York	182	173	0.56	275	0.9
OH	Ohio	143	136	0.63	274	1.1
OK	Oklahoma	148	142	0.67	258	1.0
OR	Oregon	119	116	0.65	269	1.4
PA	Pennsylvania	110	104	0.62	272	1.1
SC	South Carolina	107	104	0.24	305	1.1
TN	Tennessee	112	111	0.88	230	1.6
TX	Texas	278	270	0.61	273	0.8
WI	Wisconsin	118	117	0.75	263	1.4
WV	West Virginia	110	107	0.71	253	1.1
WY	Wyoming	80	77	0.37	293	0.9

Note. NAEP mathematics cut scores at Grade 8 are 262 for *Basic* and 299 for *Proficient.* The following states' Grade 8 mathematics test data were not used in the analysis or received special treatment: ME, MT and VA—discrepancies exist between the state assessment data and the state document; CA, KS, MN, NV, and WA—data not available in the state assessment file; AL, NH, RI, SD, UT, and VT—state assessment data not available; NE—state results are based on assessments developed by each local education agency. SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005 Mathematics Assessment, and National Longitudinal School-Level State Assessment Score Database (NLSLSASD).

[a] The proportion of NAEP sample schools employed in the estimation was less than 0.9.

Table B8

Results of Mapping State Standards to the Grade 8 NAEP Mathematics Scale: 2003

State	State name	# of schools in mapping	Estimated proportion meeting state proficiency standard, $\hat{\bar{p}}_w^B$	Estimated NAEP equivalent to state standard, $\hat{\xi}^B$	Estimated standard error of NAEP equivalent, $se\left(\hat{\xi}^B\right)$
AK	Alaska	57	0.65	268	1.5
AR	Arkansas	99	0.22	296	1.5
AZ	Arizona	105	0.21	300	1.3
CO	Colorado	104	0.68	268	1.5
CT	Connecticut	102	0.77	258	1.6
DC	District of Columbia	27	0.43	250	0.9
DE	Delaware	32	0.48	278	1.0
FL	Florida	96	0.54	269	1.7
GA	Georgia	113	0.66	255	1.3
HI	Hawaii	54	0.17	299	1.9
IA	Iowa	115	0.72	266	1.3
ID	Idaho	86	0.52	280	0.9
IL	Illinois	169	0.54	276	1.4
IN	Indiana	99	0.66	269	1.5
KY	Kentucky	112	0.32	291	1.2
LA	Louisiana	94	0.52	265	1.4
MA	Massachusetts	128	0.38	299	0.8
MD	Maryland	95	0.43	286	1.2
ME	Maine	105	0.17	311	1.0
MI	Michigan	105	0.51	278	1.4
MO	Missouri	113	0.13	314	1.0
MS	Mississippi	102	0.46	261	1.0
MT	Montana	101	0.70	271	1.0
NC	North Carolina	129	0.82	247	2.1
ND	North Dakota	31	0.44	293	1.1
NJ	New Jersey	107	0.56	278	1.3
NY	New York	141	0.54	279	1.4
OK	Oklahoma	123	0.71	256	1.5
OR	Oregon	103	0.58	275	1.6
PA	Pennsylvania	100	0.52	279	1.4
SC	South Carolina	92	0.20	306	1.5
TX	Texas	142	0.71	260	1.2
WI	Wisconsin	103	0.76	261	1.6
WY	Wyoming	74	0.35	297	1.1

Note. NAEP mathematics cut scores at Grade 8 are 262 for *Basic* and 299 for *Proficient.* (Median SE of the NAEP equivalent = 1.4.) Results were deleted for VA due to discrepancies between state assessment data and the state document. SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2003 Mathematics Assessment, and National Longitudinal School-Level State Assessment Score Database (NLSLSASD).